Intergrating LLM-Based Indicators of Compromise Extraction in Cybercrime Intelligence Analysis

Tim Angevare
University of Twente
t.p.angevare@student.utwente.nl
s2744007

ABSTRACT

Research by Grochmal et al. on the challenges faced by law enforcement found that modern devices and cloud storage confiscated in cybercrime investigations generate data volumes that exceed the analytical and storage capacity of many agencies [1]. These bottlenecks limit capacity and resources in later phases of the investigation. Automated Indicator of Compromise (IoC) extraction from unstructured text typically relies on pattern-matching techniques such as regular expressions or other Natural Language Procession (NLP) algorithms, which struggle with diverse formats and contextual variations inherent in cybercrime data. While existing research explores LLM-based entity extraction, current approaches fail to address law enforcement-specific operational constraints including regulatory compliance and practical on-premises deployment requirements, nor do they systematically compare multiple open-source LLMs or integrate them with traditional NLP methods. This research systematically evaluates approximately 20 open-source LLMs for cybercrime IoC extraction through prompt engineering and Low-Rank Adaptation (LoRA) fine-tuning. Subsequently, it designs an integration methodology that leverages the strengths of both LLM-based and pattern-matching approaches for optimal extraction, transformation, and doading (ETL) automation. Target entity types include IP addresses (v4/v6), email addresses, domain names, usernames, phone numbers, and financial accounts including cryptocurrency addresses. This work delivers the first systematic LLM methodology specifically validated for law enforcement IoC extraction contexts and provides a validated integration approach for transforming manual ETL processes into automated, compliance-ready extraction pipelines, addressing a critical operational bottleneck in the cybercrime intelligence cycle.

1. INTRODUCTION

Investigative law enforcement agencies, such as Europol, investigate subjects and events after suspicion of illegal activities. Law enforcement investigations rely fundamentally on the initiation of the project with the intelligence cycle. The process of an investigation begins with the collection of data on or related to the case from various origins. Europol, cross-references data obtained from national law enforcement agencies for analysis in order to coordinate the international investigation in the latter dissemination phases. A crucial part of the initial analysis for law inforcement is extracting forensic artifacts that are signs that a system has been compromised by an attack or that it has been infected with a particular malicious software. These are called Indicators of Compromise (IoCs). Examples of IoCs include entities such as IP addresses (v4 or v6), email addresses, domain names, usernames, cryptocurrency addresses, organizations, names, nicknames and locations. IoCs can reveal information on threat actor infrastructure, link threat actors in multiple cases, correlate evidence, identify victims, or serve as tactical intelligence for infrastructure seizure.

Due to the ongoing digital transformation of society and the nature of cybercrime investigations, the amount, format, and complexity of data collected by law enforcement is exceeding analytical and storage capacity. Automated IoC extraction from unstructured text typically relies on pattern-matching techniques such as regular expressions (regex) and rule-based Natural Language Processing (NLP) methods. However, these approaches struggle with the diverse formats and contextual variations inherent in law enforcement cybercrime data. Consequently, extraction, transformation, and loading (ETL) of investigative data remains largely manual, repetitive work. For example, the same phone number can be written in the following accepted formats:

- 0703561220 (without formatting)
- 070 356 12 20 (spaces)
- 070-356 12 20 (dash after the area code)
- (070) 356 12 20 (parentheses around the area code)
- +31 70 356 12 20 (international notation)

Or a combination of multiple. In addition, the format will also change depending on the source of the data. For Hyper Text Markup Language (HTML), typically used for web pages, it can be: `<a href="tel:+31703561220">`.

In other cases, subjects might deliberately try to obfuscate IoCs to evade extraction by using methods that make sense to a human analyst but are missed by RegEx or NLP methods. An example of this is using the email: example[at]domain.com

This research first explores the current state-of-the art IoC extraction methods applicable for unstructured cybercrime texts. Subsequently, it researches how LLMs can be utilized to achieve competitive or superior performance in cybercrime contexts. Finally this research designs a tool that streamlines the ETL pipeline for IoC extraction in unstructured cybercrime texts by integrating LLM-based entity extraction with pattern-matching techniques. To pursue our goal, we have defined the following research questions (RQ) as the basis of our research:

- **RQ1:** What are the current state-of-the art NER methods for extracting IoC from unstructured cybercrime intelligence texts?
- **RQ2:** How can LLMs achieve competitive or superior performance in terms of precision, recall and F1-scores compared to state-of-the-art NER methods for IoC extraction?
- **RQ3:** How can we leverage the strengths of and integrate the current-state-of-the art methods along with LLMs for optimal automatization of extracting IoCs in terms of time reduction and accuracy?

The results of each research question will serve as the input for the subsequent questions. The remainder of this research proposal is organized as follows. The next section will present the related work and argue the gap in the research. Then, section 3 will discuss the approaches expected for answering each research question. The proposal is concluded with a preliminary planning for the research questions and an overview of the milestones shown in Section 4.

This research is expected to contribute a validated systematic methodology for LLM-based IoC extraction in unstructured cybercrime texts for operational law enforcement contexts. In addition, this research will contribute a design that integrates LLMs with pattern-matching techniques for reducing manual ETL workload and time while retaining a high accuracy. This work addresses a critical bottleneck in the cybercrime intelligence cycle for law enforcements.

2. RELATED WORK

Natural Language Processing is a broad domain and with that already a substantial amount of research has been done in IoC extraction. Some of the more recent papers have looked at the applicability of machine learning and LLMs in this field. Nevertheless, to the best of my knowledge there is no research that compares the performance of multiple open-source LLMs and further explores fune-tuning techniques for IoC extraction unstructured texts for law enforcement contexts.

Paper	Similarities	Difference
Entity extraction of key elements in 110 police reports based on large language models [2]	Uses open source large language models for entity extraction. Extraction of descriptive elements such as organizations, actions, locations.	Only focuses on entity extraction and on police reports on a certain subject. Focuses on key elements describing a situation and not on personal data. The data is in mandarin.
Evaluating LLM-based Personal Information Extraction and Countermeasures [3]	Researches Personal Information Extraction using LLMs.	Does not focus on events and researches defenses and defense bypasses which are out of scope for this research. Extraction from public profiles which can be a part of the dataset but is not entirely. Most of the dataset is in HTML format.
Leveraging Open Large Language Models for Historical Named Entity Recognition [4]	Researches LLMs for NER in historical texts.	For applications of large historical texts and does not go into relation extraction.
Privacy BERT-LSTM: a novel NLP algorithm for sensitive information detection in textual documents [5]	Use of deep learning in order to extract sensitive personal data in texts.	Researches BERT models instead of LLMs and does not look into relations.
IoCMiner: Automatic Extraction of Indicators of Compromise from Twitter [6]	Extraction of IoCs from unstructured primary sources .	Uses a combination of machine learning classification and regular expressions.

Table 1: Related Work

3. METHODOLOGIES

The research will follow a mixed-methods approach with a combination of literature review, experimental evaluation, and design science research for the tool development in RQ3.

3.1 On answering RQ1

A systematic review of the literature will be done to identify and analyze the current state-of-the-art IoC extraction methods. The strengths and shortcomings will be weighed and this framework will be used as a baseline for the evaluation of the LLMs in RQ2. The answer of RQ1 is the gap in research that justifies the existence of RQ2.

3.2 On answering RQ2

In RQ2 quantitative systematic experiments will be conducted to evaluate the performance of LLMs in IoC

extraction compared to state-of-the-art method.

First, the dataset will be composed of real anonymized or synthetic data generated with LLMs that reflect on real patterns validated by Europol experts to ensure representativeness of operational scenarios without compromising sensitive data in combination with leaked cybercrime chats. Target entity types will include IP addresses (v4 and v6), names, nicknames, locations, email addresses, web resources, usernames, phone numbers, and financial accounts such as back accounts and cryptocurrency addresses.

A set of approximately 20 open-source LLMs across different model families and parameter sizes will be evaluated against the dataset with a task and format specification prompt in order to establish a baseline performance. This experiment will be conducted on a GPU cluster from the University of Twente. Model performance is evaluated using precision, recall, F1-score, and elapsed time. The performance will then be improved through two prompt engineering strategies: (1) role, task, and format specification and (2) role, task, format and few-shot examples. Each prompt strategy will be evaluated against 5 test samples across all models, resulting in 300 experiment iterations.

For the performing model, further LoRA fine-tuning will be conducted in addition with context tuning of the text input for optimal performance. The final performance will be bench-marked against the current state-of-the-art methods and the model will serve as the input for the tool development in RQ3.

3.3 On answering RQ3

This research question will focus on the design of a proof-of-concept tool to integrate current state-of-the-art methods and LLMs in order to leverage the strengths of both methods of for optimal performance in terms of accuracy and time reduction. The optimal integration will be researched by systematically testing various IoC extraction methods, sequences, and split on the dataset while recording metrics such as precision, recall, F1-score, and elapsed time. We will compare the results and argue which divisions are best in which cases.

The methodology for the design of the tool will adhere to the design research paradigm and focus on the following steps:

1. **Problem identification and motivation:** Like discussed in section 1 of this proposal.
2. **Objective of a solution:** Done through interviews resulting in a list of requirements.
3. **Design and Development:** Creating the artifact. We specify system requirements based on expert interviews and Europol's intelligence cycle process in order to finally further generalise to the sector. With the requirements and the constraints the tool architecture will be defined.
4. **Demonstration:** Using the artifact to solve a problem. The tool will be demonstrated to Europol's EC3 unit analysts.
5. **Evaluation:** Measuring how well the artifact achieves its objectives. The tool will be evaluated against the dataset in RQ2 along with expert interviews with Europol analysts in order to validate its practical use and shortcomings.

of The Process Model (Peffers et al., 2020) [7]

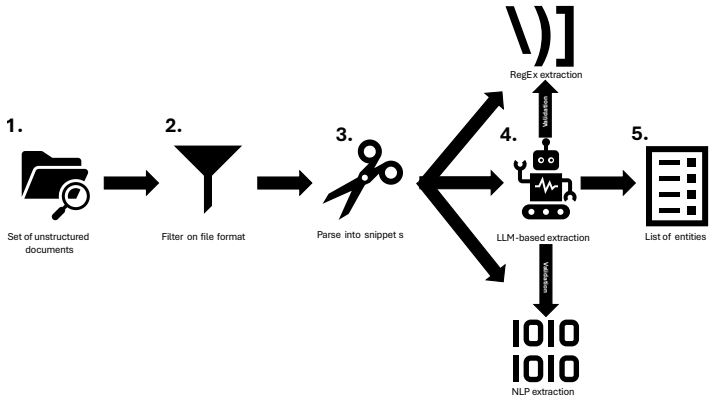


Figure 1: Proposed process with implementation of tool

4. PLANNING

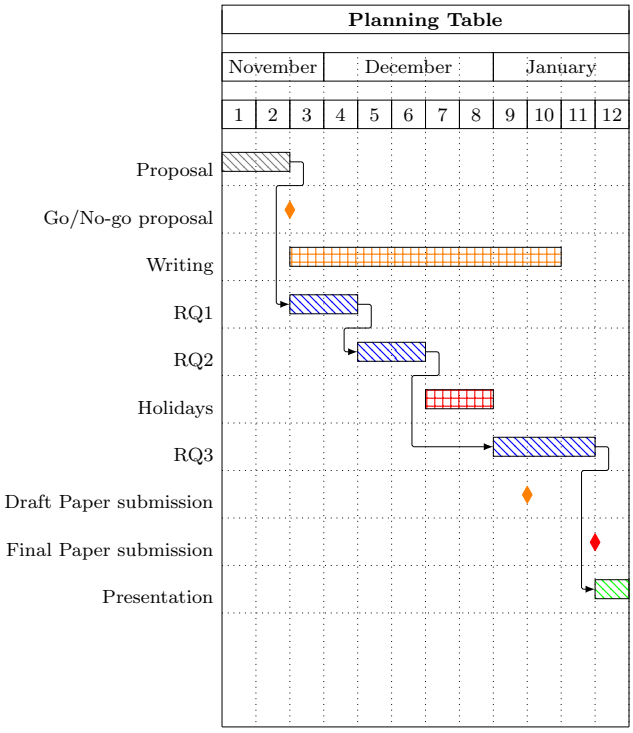
This section will discuss the planning of the research and show an overview of the milestones. The study has been split into four parts, as can be seen in the table below. Note that a planning such as this is to be seen as a guideline. There are, however, some hard deadlines for handing in drafts and final versions. Here we have an overview of the deadlines:

- November 23rd: Go/no-go proposal submission
- January 18th: Draft paper submission
- January 25th: Final paper submission
- January 30th: Conference presentation

The planning is made in order to adhere to these submission deadlines.

References

Alexis B. Grochmal. *Challenges Faced by Law Enforcement Collecting and Using Digital Evidence in Cybercrime Investigations.* PhD thesis, Marymount University, 2025. URL https:



//www.proquest.com/dissertations-theses/challenges-faced-law-enforcement-collecting-using/docview/3193985896/se-2.

Xintao Xing and Peng Chen. Entity extraction of key elements in 110 police reports based on large language models. *Applied Sciences*, 14(17), 2024. ISSN 2076-3417. doi: 10.3390/app14177819. URL https://www.mdpi.com/2076-3417/14/17/7819.

Yupei Liu, Yuqi Jia, Jinyuan Jia, and Neil Zhenqiang Gong. Evaluating llm-based personal information extraction and countermeasures. In *USENIX Security Symposium*. USENIX Association, 2025.

Carlos-Emiliano González-Gallardo, Hanh Thi Hong Tran, Ahmed Hamdi, and Antoine Doucet. Leveraging open large language models for historical named entity recognition. In Apostolos Antonacopoulos, Annika Hinze, Benjamin Piwowarski, Mickaël Coustaty, Giorgio Maria Di Nunzio, Francesco Gelati, and Nicholas Vanderschantz, editors, *Linking Theory and Practice of Digital Libraries*, pages 379–395, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72437-4.

Janani Muralitharan and Chandrasekar Arumugam. Privacy bert-lstm: a novel nlp algorithm for sensitive information detection in textual documents. *Neural Computing and Applications*, 36(25):15439–15454, May 2024. doi: 10.1007/s00521-024-09707-w. URL https://doi.org/10.1007/s00521-024-09707-w.

Amirreza Niakanlahiji, Lida Safarnejad, Reginald Harper, and Bei-Tseng Chu. Iocminer: Automatic extraction of indicators of compromise from twitter. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4747–4754, 2019. doi: 10.1109/BigData47090.2019.9006562.

Ken Peffers, Tuure Tuunanen, Charles E Gengler, Matti Rossi, Wendy Hui, Ville Virtanen, and Johanna Bragge. Design science research process: A model for producing and presenting information systems research, 2020. URL https://arxiv.org/abs/2006.02763.